AJ 156, 58 (2018)
Typeset using LaTeX default style in AASTeX62

ESTIMATING DISTANCES FROM PARALLAXES IV: DISTANCES TO 1.33 BILLION STARS IN *Gaia* DATA RELEASE 2

C.A.L. Bailer-Jones,[1] J. Rybizki,[1] M. Fouesneau,[1] G. Mantelet,[2] and R. Andrae[1]

[1] *Max Planck Institute for Astronomy, Heidelberg, Germany*
[2] *Astronomisches Rechen-Institut, Zentrum für Astronomie der Universität Heidelberg, Germany*

(Received 26 April 2018; Revised 30 May 2018; Accepted 5 June 2018; Published 20 July 2018)

ABSTRACT

For the vast majority of stars in the second *Gaia* data release, reliable distances cannot be obtained by inverting the parallax. A correct inference procedure must instead be used to account for the nonlinearity of the transformation and the asymmetry of the resulting probability distribution. Here we infer distances to essentially all 1.33 billion stars with parallaxes published in the second *Gaia* data release. This is done using a weak distance prior that varies smoothly as a function of Galactic longitude and latitude according to a Galaxy model. The irreducible uncertainty in the distance estimate is characterized by the lower and upper bounds of an asymmetric confidence interval. Although more precise distances can be estimated for a subset of the stars using additional data (such as photometry), our goal is to provide purely geometric distance estimates, independent of assumptions about the physical properties of, or interstellar extinction towards, individual stars. We analyse the characteristics of the catalogue and validate it using clusters. The catalogue can be queried on the Gaia archive using ADQL at http://gea.esac.esa.int/archive/ and downloaded from http://www.mpia.de/~calj/gdr2_distances.html.

1. INTRODUCTION

Essentially all published quantities are inferences as opposed to direct measurements. Even apparently simple quantities such as celestial coordinates are achieved only after extensive data processing, e.g. reducing CCD images, fitting a point spread function, applying a focal plane geometric calibration, transforming to a global coordinate system. The role and impact of the assumptions involved in this process should not be overlooked.

The parallaxes published in the second *Gaia* data release (hereafter GDR2; Gaia Collaboration 2018a) are obtained after a complex iterative procedure, involving various assumptions (Lindegren et al. 2012). Going from a parallax to a distance is also a non-trivial issue, as has been described in several previous works (see Luri et al. 2018 for a recent discussion). The main issues are the nonlinearity of the transformation and the positivity constraint of distance (but not the parallax). Many parallaxes in GDR2 also have very low signal-to-noise ratios (this will be quantified later), meaning that a characterization of the inferred distance uncertainties is at least as important as a single point estimate of the distance.

The only consistent and physically meaningful way to infer distances and their uncertainties is through a probabilistic analysis, as described for example in Bailer-Jones (2015). This involves specifying a likelihood and a prior. The likelihood is defined by the parallax inference process, and on account of the assumptions involved, the linearity of the model, and the central limit theorem, this is theoretically a Gaussian distribution (L. Lindegren, private communication), and it has also been confirmed empirically that this is a very good approximation for GDR2 (Lindegren et al. 2018). The prior should incorporate any relevant aspects of the expected distance distribution that are not contained within the likelihood. The obvious one is the positivity of distance. Others might be the properties of the survey and our current knowledge of the Galaxy.

Corresponding author: Coryn A.L. Bailer-Jones
calj@mpia.de

What kind of prior one adopts depends in part on the desired trade-off between model-dependence and precision: a complex prior based on a detailed Galaxy model will follow this more closely in the limit of poor data, which may or may not be what one wants. As such one should not single out priors for introducing "biases", as their use is conceptually no different from what is done in any inferential process, which always involves a combination of both "measurements" and assumptions. Indeed, there is no such thing as uninformative priors or model-free inference. The issue is not whether a bias is present, but what impact our choices have. The advantage of the probabilistic approach is that it combines the measurement (likelihood) and assumptions (priors) in a consistent way that makes the prior irrelevant when the data are good, but ensures a graceful transition to dominance by the prior as the data quality degrades.

Bailer-Jones (2015) and Astraatmadja & Bailer-Jones (2016a) (papers I and II in this series) investigate the consequences of using different types of priors, including one based on a Galaxy model. Using the first *Gaia* data release of two million parallaxes, Astraatmadja & Bailer-Jones (2016b) (paper III) inferred distances using priors at two extremes: (1) a simple isotropic prior; (2) a prior given by the distribution of stars along each line-of-sight as determined from a Galaxy model, which also accounted for interstellar extinction and the *Gaia* selection function.

In this paper we infer distances to the 1.33 billion sources in the GDR2 catalogue that have parallaxes. We adopt the exponentially decreasing space density (EDSD) prior in distance r

$$P(r \mid L) = \begin{cases} \frac{1}{2L^3} r^2 e^{-r/L} & \text{if } r > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $L > 0$ is a length scale, as described in Bailer-Jones (2015) and Astraatmadja & Bailer-Jones (2016a), and adopted by Astraatmadja & Bailer-Jones (2016b) for GDR1. This prior has a single mode at $2L$. In the current paper L varies as a function of longitude and latitude (l, b) according to a model, to reflect the expected variation in the distribution of stellar distances in the *Gaia*-observed Galaxy. This is, in some sense, a compromise between the simplicity of the isotropic prior (i.e. fixed L) and the complexity of the line-of-sight-dependent distribution shapes obtained from a Galaxy model.

In this work we only use the *Gaia* astrometry and our length scale model to infer distances. Given the limited fractional precision of the parallaxes for more distant stars, this limits the distance precision we can achieve. One could, of course, use other information to improve the distance estimates further. An obvious addition is to use a model for the absolute magnitude of the star together with a measurement of its apparent magnitude and a measurement of (or model for) the line-of-sight extinction. An absolute magnitude model is normally based on the Hertzsprung–Russell diagram and the observed colours or spectroscopy of the star (e.g. Astraatmadja & Bailer-Jones 2016a). Such an approach was taken by McMillan et al. (2017) to estimate distances using the overlap between RAVE spectroscopy and the two million parallaxes in GDR1. Other recent studies combining parallaxes with other information to improve distance measures include Schönrich & Bergemann (2014), Leistedt & Hogg (2017), O'Malley et al. (2017), Anderson et al. (2017), Queiroz et al. (2018), Cantat-Gaudin et al. (2018), Coronado et al. (2018), and Fouesneau et al. (in preparation).

Our goal here is different. We choose to infer distances without invoking any assumptions about the properties of, or the extinction towards, individual stars. (Stellar properties and extinction only enter in a collective sense to construct the prior.) This enables us to produce a self-consistent catalogue for all of GDR2. The price we pay is precision. Our catalogue provides a purely geometric measure of distance and its uncertainty that overcomes the dangers of inverse parallax and unphysical priors. The obvious use case for this catalogue is to provide the distance to one or more stars, or rather the probable range of distances to the stars, since the uncertainties are real and often significant. This can tell us, for example, whether a set of stars have consistent distances. The catalogue can be used to select a sample of stars on which other inferences are then performed, or for which more precise distances are estimated using additional information. The catalogue distances also serve as a baseline against which to compare other distance estimates. We may likewise use the catalogue to determine the three-dimensional space distribution of a set of stars. Note that although the distances are determined independently from one another, the prior is correlated on small spatial scales. Thus if we know or suspect that the stars are members of a stellar cluster, a combination of our distances may not provide a reliable distance to the cluster. It is far better to set up a model for the cluster in which its distance is a free parameter, and to solve for this using the original parallaxes (accommodating also their spatial correlations). Generally speaking it is suboptimal to use our distances if they are just an intermediate step in a calculation, e.g. if

one really wants to estimate absolute magnitudes or the transverse velocity of a cluster. More accurate results (and a better propagation of uncertainties) will be obtained by inferring directly the quantities of interest.

In the next section we describe the prior model and the computation of distances and asymmetric uncertainties. In section 3 we analyse the results. The content and characteristics of the catalogue are presented in section 4. We summarize in section 5 with some more notes on how (not) to use the catalogue.

2. METHOD

2.1. *Posterior probability density function*

Given the Gaussian likelihood in the parallax ϖ with standard deviation σ_ϖ, plus the EDSD prior from equation 1, the unnormalized posterior over the distance to a source is

$$P^*(r \mid \varpi, \sigma_\varpi, L_{\rm sph}(l,b)) \;=\; \begin{cases} r^2 \exp\left[-\dfrac{r}{L_{\rm sph}(l,b)} - \dfrac{1}{2\sigma_\varpi^2}\left(\varpi - \varpi_{\rm zp} - \dfrac{1}{r}\right)^2\right] & \text{if} \quad r > 0 \\ 0 & \text{otherwise} \,. \end{cases} \tag{2}$$

The quantities ϖ, σ_ϖ, l, b are all taken from the `gaia_source` table in GDR2 from the columns `parallax`, `parallax_error`, `l`, `b` respectively. $\varpi_{\rm zp}$ is the global parallax zeropoint, determined from *Gaia*'s observations of quasars to be -0.029 mas (Lindegren et al. 2018). (The zeropoint actually varies as a function of sky position, magnitude, and colour, but this is the best global estimate.) The length scale $L_{\rm sph}(l,b)$ has a Galactic longitude and latitude dependence as described in section 2.3.

The properties of this posterior have been explored in some detail in Bailer-Jones (2015). For physical values of its three parameters – finite ϖ, positive σ_ϖ, positive $L_{\rm sph}$ – it is always a proper (i.e. normalizable) density function. It can be either unimodal or bimodal, although for the vast majority of cases in GDR2 it is unimodal.

2.2. *Distance estimators*

While the posterior in equation 2 is the complete description of the distance to the source (fully specified by ϖ, σ_ϖ in GDR2 and $L_{\rm sph}$ in our catalogue), we sometimes want a point estimate along with some measure of the uncertainty. As known from previous work and seen from the examples in section 3, this posterior is asymmetric, with a longer tail towards large distances. Being a proper (normalizable) posterior, all of the usual estimators – mean, median, mode, variance, quantiles, full-width at half-maximum – are defined. As the point estimator, $r_{\rm est}$, we prefer here the mode, $r_{\rm mode}$. This is found analytically by solving a cubic equation (Bailer-Jones 2015).

As a measure of uncertainty we use the highest density interval (HDI) with probability p. The HDI is the span of the distance that encloses the region of highest posterior probability density, the integral of which is p. The span is defined by the lower and upper bounds, $r_{\rm lo}$ and $r_{\rm hi}$ respectively. Here we set $p = 0.6827$, which is equal to the probability contained within $\pm 1\sigma$ of the mode for a Gaussian distribution. The distance posterior is asymmetric (sometimes significantly so) so $r_{\rm hi} - r_{\rm est}$ and $r_{\rm est} - r_{\rm lo}$ are unequal. Conceptually, the HDI can be found by lowering a horizontal line over the distribution until the area contained under the curve between its intercepts with the curve (i.e. $r_{\rm lo}$ and $r_{\rm hi}$) is equal to p (see Figure 5.10 of Bailer-Jones 2017). The HDI has the advantage over other confidence intervals in that it is guaranteed to contain the mode. Some other common measures are less flexible. The Fisher information approach, for example, assumes the posterior is locally Gaussian, the standard deviation of which is taken as the uncertainty (which is symmetric and therefore entirely inappropriate for positively-constrained quantities like distances).

This HDI is unique if the posterior is unimodal. (A more general way of finding the HDI allows it to split into multiple intervals in the case of multimodality; see for example Hyndman 1996). For a small part of the parameter space defined by $(\varpi, \sigma_\varpi, L)$, the posterior is bimodal (see Bailer-Jones 2015). For these cases we retain both the mode estimator and the HDI if possible (i.e. if the span does not include the minimum, thereby retaining the uniqueness; see below). Otherwise we resort to using the median of the distribution as the point estimator, and report the 16th and 84th percentiles (i.e. $(1 \pm p)/2$) as $r_{\rm lo}$ and $r_{\rm hi}$ respectively; together these two numbers form the equal-tailed interval (ETI), which has as much probability below the span as above, with p in between.

There is no analytic solution for the HDI for this posterior. We compute it with an iterative procedure involving a Taylor expansion. The basic idea is to take a small step in both directions away from the mode, compute the area under the curve covered by this step, and iterate this until the total area hits the desired limit.

We first normalize the posterior using Gaussian quadrature; denote this with $P(r)$. With r set equal to the mode (where the first derivative is zero), and adopting a fixed negative ΔP, an initial step of size

$$\Delta r_0 = \sqrt{2\Delta P \left(\frac{\mathrm{d}^2 P}{\mathrm{d}r^2}\right)^{-1}_{r_{\mathrm{mode}}}} \tag{3}$$

is computed. Candidate lower and upper bounds are then defined as

$$r_1^{\pm} = r_{\mathrm{mode}} \pm \Delta r_0 \ . \tag{4}$$

The area under the curve between these bounds is computed using the trapezium rule. Assuming this area is less than p (ΔP is set small enough to ensure this is always the case), further steps are computed in both the negative and positive directions using the first order term in the Taylor expansion, i.e.

$$\Delta r_i^{\pm} = \Delta P \left(\frac{\mathrm{d}P}{\mathrm{d}r}\right)^{-1}_{r_i^{\pm}} \tag{5}$$

starting from $i = 1$, to give new candidate lower and upper bounds as

$$r_{i+1}^{\pm} = r_i^{\pm} + \Delta r^{\pm} \ . \tag{6}$$

The additional area covered by these steps is estimated using the trapezium rule and added to the running total area. This process is iterated from equation 5, with the result that both bounds move monotonically away from the mode. The procedure is stopped when the running total area computed exceeds p (or bimodality is detected; see below). For the distance catalogue produced here, ΔP is fixed to $-0.01P(r_{\mathrm{mode}})$. For 98% of the sources in GDR2, between 39 and 73 iterations (the 1st and 99th percentiles) are required.

Note that the area computed by this method actually exceeds p, although generally by only a very small amount: for 99.9% of the sources the area computed is still less than 0.6973. This is an acceptable degree of approximation. (Tests shows that the area calculation by the trapezium rule is generally very accurate: less than 0.06% of cases deviate from an accurately-calculated area by more than 1 part in 1000.) In a few pathological cases the posterior has a long, nearly flat tail extending to large distances. The numerical estimation of the HDI is then poor. In a very small fraction of cases this can even lead to the computed area exceeding 0.9. If this happens the mode and HDI are rejected as catalogue entries, and replaced with the median and equal-tailed quantiles, as explained below. (These solutions have `result_flag=2` and `modality_flag=1`. A full description of the flags is given in the next section.)

Because the steps in the positive and negative directions are made independently, the values of the posterior at r_{lo} and r_{hi} are not identical: an imaginary line being lowered over the posterior does not remain exactly horizontal. Because $P = 0$ only for $r = 0$ and $r = +\infty$ (equation 2), and because r_{hi} can never reach infinity, then in principle r_{lo} should never reach exactly zero. With this numerical method it can, although it only happens in 183 765 (0.014%) cases. The algorithm prevents steps to negative distances, so $r_{\mathrm{lo}} \geq 0$. (Note that this method of finding the HDI would not work without modification for distributions which drop to zero at finite values of r greater than zero.)

If the posterior is bimodal (which occurs in 0.09% of cases), then the search for the HDI is done starting from the highest mode. Provided the search does not encounter the minimum between the two modes, the HDI remains well defined, and this, together with the highest mode, are the posterior summary provided in the catalogue. The `result_flag` has value `1` whenever the point estimate, r_{est}, is the mode, and the lower and upper bounds of the confidence interval, r_{lo} and r_{hi}, define the HDI.

The presence of bimodality (which is identified independently of the search, from the roots of $\mathrm{d}P/\mathrm{d}r = 0$) is always indicated by the `modality_flag`, which is `1` for unimodal and `2` for bimodal. The minimum is encountered during the HDI search in less than 0.1% of all cases. If this happens we stop the search and instead compute the median (the 50th percentile) and the ETI, which we report in the catalogue. Such solutions are indicated with `result_flag=2`. The quantiles are found numerically by drawing 2×10^4 samples from the posterior with a Markov Chain Monte Carlo (MCMC) method. Depending on the shape of the posterior, these estimates of the quantiles may not be very accurate, so should be treated with caution.

The HDI is a couple of orders of magnitude faster to compute than the ETI, because the former involves that many fewer calculations of the posterior density.

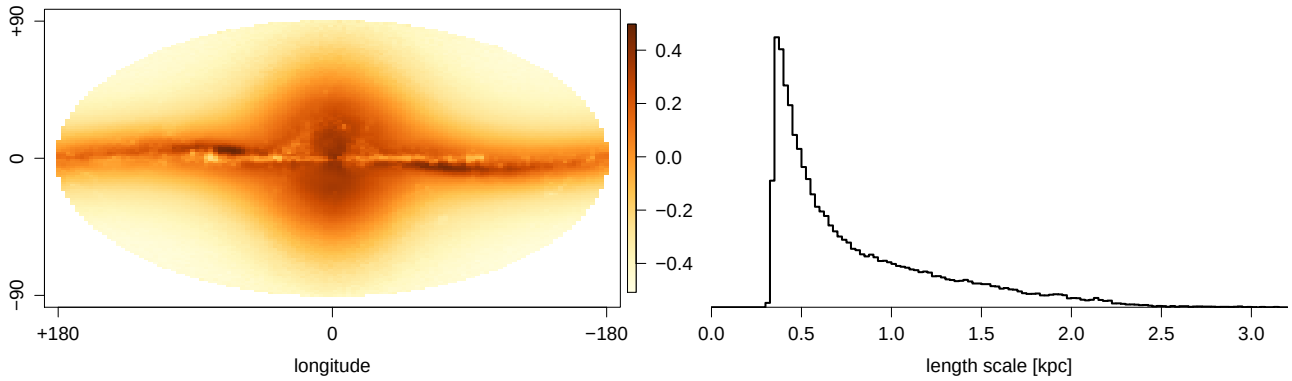

Figure 1. Distribution of the length scales from the Galaxy model, $L_{\rm gal}(l,b)$. Left: The log (base 10) length scale in kpc shown in Galactic coordinates in a Mollweide (equal-area) projection. Right: Histogram (linear units on both axes) of the length scales (over equal-area cells). The full range of $L_{\rm gal}$ in the model (and covered by the colour scale) is 0.310 to 3.143 kpc (-0.508 to 0.497 on the log scale).

In a very small number of cases (about two in every million), the algorithm fails to give any summary of the posterior. In these cases `result_flag=0` and the three distances $r_{\rm est}$, $r_{\rm lo}$, $r_{\rm hi}$ are set to `NaN`. This occurs when the parallax is negative yet highly statistically significant, $\sigma_\varpi/\varpi \ll -1$, which results in the unnormalized posterior density being numerically equal to zero everywhere on account of the finite machine precision. This precludes both the computation of the normalization constant (for the HDI) and an MCMC sampling (for the quantiles). Although we can still compute the mode (it is found algebraically), we choose not to report it. One could work around this problem by including scaling factors in the computation of the density, but the cases are very rare, plus the distance posterior is virtually identical to the prior, which is entirely specified by the length scale (always provided).

We publish only a single point estimate together with the bounds of a single confidence interval in order to keep the catalogue manageable in size. Should users want other estimates, e.g. different size confidence intervals, it is quick and easy to recompute the posterior using ϖ and σ_ϖ in GDR2 and $L_{\rm sph}$ in our catalogue. Processing all 1.33 billion sources took 75 CPU-core days, which on our multicore machine was just a few hours of wall-clock time. The code for the processing is available at https://github.com/ehalley/Gaia-DR2-distances.

2.3. *Length scale model*

We use a chemo-dynamical model to simulate the Galaxy (including extinction) as seen by *Gaia*, which here means all stars with apparent magnitude $G \leq 20.7$ mag. The resulting mock catalogue and the Galaxy model on which it is based are described in Rybizki et al. (2018). From this catalogue we extract the distances to all stars in each of the 49 152 equal-sized (0.84 square degrees) healpix cells (level 6) over the entire sky. We then fit the EDSD prior, equation 1, to the stars in each cell. The maximum likelihood estimate for L is just $\overline{r}/3$, where $\overline{r}$ is the mean of the distances in that cell. To avoid a few stars at very large distances dominating this estimate we replace the mean with the median.

The resulting map, $L_{\rm gal}(l,b)$, is shown in Figure 1. The bulge, disk, and warp of the disk are quite prominent. Perhaps counter-intuitively, the median distance is generally smaller at high Galactic latitudes than at low latitudes. This is because of the very high density of stars far away in the Galactic disk which *Gaia* can see, despite the extinction. Although the maximum distance *Gaia* can observe to is larger at high latitudes, the density of star drops rapidly with height above the disk, so the median is determined primarily by nearer stars. The same map computed using $\overline{r}/3$ exhibits larger distances than this map at high latitudes.

We do not use this map directly as the prior model, because it shows discontinuities across the healpix boundaries. This would lead to spatial discontinuities in the inferred distances. We therefore fit a spherical harmonic model of order $n_{\rm max}$

$$f(l,b)_{n_{\rm max}} \;=\; a + \sum_{n=1}^{n_{\rm max}} \sum_{m=0}^{n} s_{n,m} P_n^m(\sin b)\sin(ml) + c_{n,m} P_n^m(\sin b)\cos(ml) \tag{7}$$

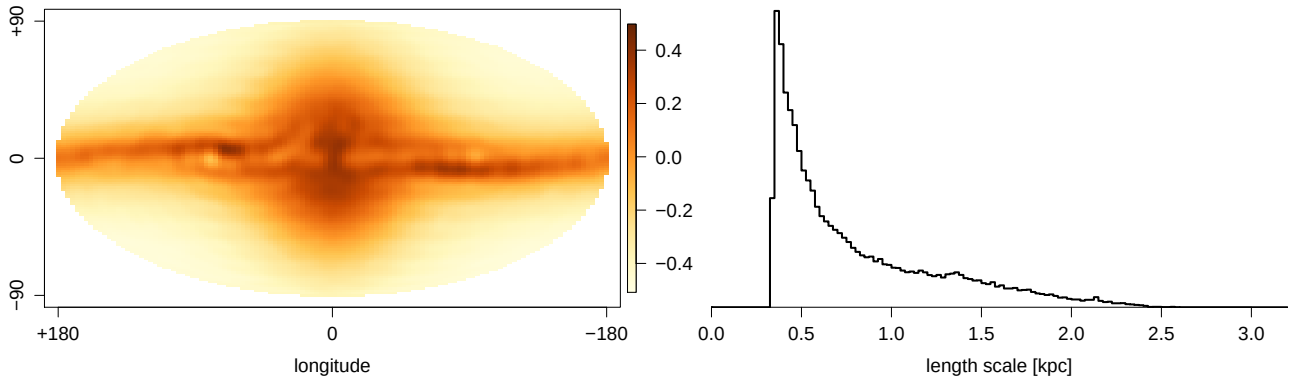

Figure 2. Distribution of the spherical harmonic model fit to the length scales, $L_{\rm sph}(l,b)$. Left: The log (base 10) length scale in kpc shown in Galactic coordinates in a Mollweide projection. Apparent discontinuities in the map are due to the finite-sized cells used for plotting; they are not a property of the fit. Right: Histogram (linear units on both axes) of the length scales (over equal-area cells). The plotting ranges of the distance in both panels (colour scale and histogram) are the same as in Figure 1. The range of $L_{\rm sph}$ values in the catalogue is given in Table 2.

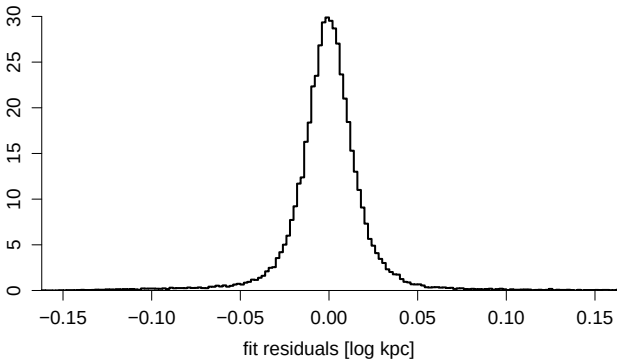

Figure 3. Histogram of the residuals of the fit of the spherical harmonic model, in the sense of $\log_{10} L_{\rm sph} - \log_{10} L_{\rm gal}$.

where $P_n^m()$ is the associated Legendre polynomial of order (n,m), and $\{s_{n,m}, c_{n,m}\}$ together with a are the free parameters to be determined. Note as $P_n^m(\sin b)\sin(ml) = 0$ when $m = 0$, the constants $s_{n,0}$ are unconstrained, so we set them to zero. The total number of free parameters is

$$1 + \sum_{n=1}^{n_{\rm max}} (2n+1) = (n_{\rm max}+1)^2. \tag{8}$$

We fit this model to the map of $\log_{10} L$ using linear least squares. After some experimentation, also with the healpix level of the input map, we settled on $n_{\rm max} = 30$ as high enough to capture the essential spatial variations, but low enough to smooth out some of the rather sharp features in the model. The fitted map, which we denote $L_{\rm sph}(l,b)$, is shown in Figure 2. A histogram of the residuals of the fit is shown in Figure 3.

3. ANALYSIS OF DISTANCE RESULTS

We now analyse the results of the distance inference, using a random subset of a million sources.

3.1. *Parallax statistics*

The distribution of the parallaxes is shown in Figure 4 (compare with the predictions in Figure 2 of Astraatmadja & Bailer-Jones 2016a). 24.9% of the sources have negative parallaxes. The distribution of the fractional parallax uncertainty, defined as σ_ϖ/ϖ, is shown in panel (b). Immediately striking is how many sources have very low signal-to-noise (SNR) parallaxes: of those sources with positive parallaxes, only 15.7% have $\sigma_\varpi/\varpi < 0.2$, i.e. have a SNR

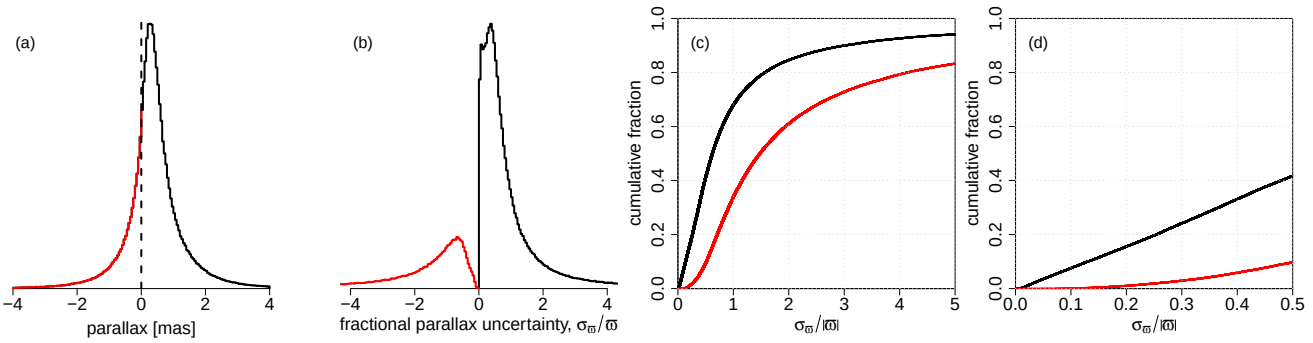

Figure 4. The parallaxes and parallax precisions in GDR2. (a) The distribution of parallaxes in GDR2. (b) The distribution of fractional parallax uncertainties, σ_ϖ/ϖ. Linear scales are used on both axes of the histograms. This is shown as a cumulative distribution in panel (c) split for the positive parallaxes (black/upper line) and the negative parallaxes (red/lower line) (i.e. the two lines asymptote to values which sum to 1.0). Panel (d) is a zoom of (c).

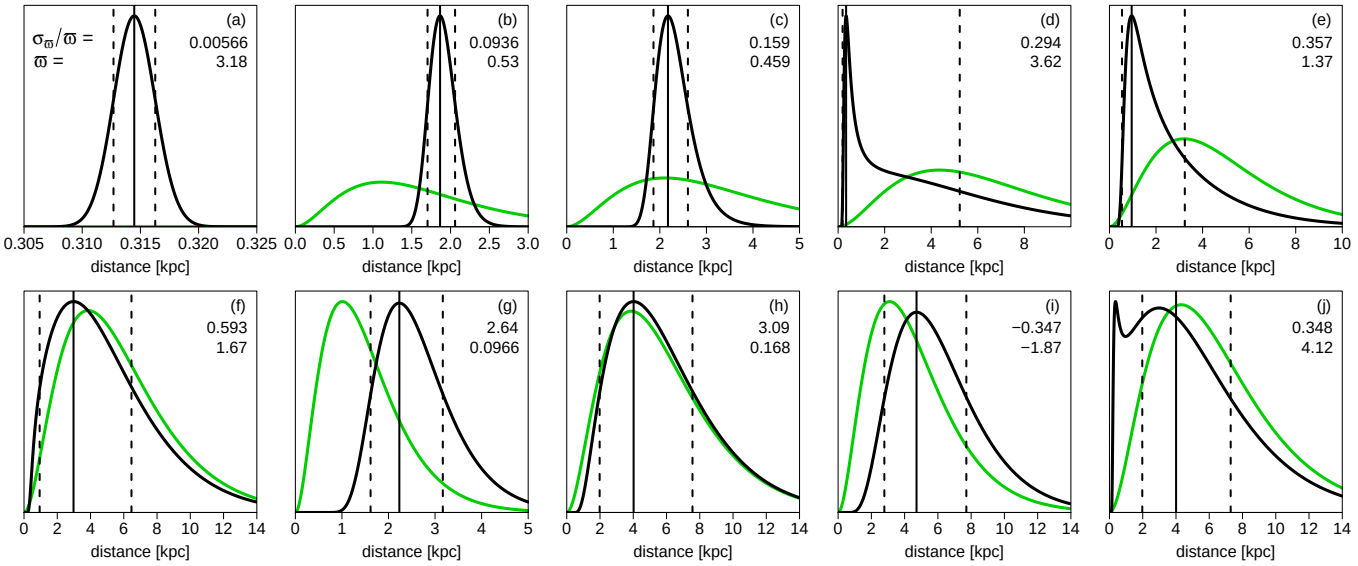

Figure 5. Ten examples of the inferred posteriors (black curves) and the corresponding priors (green curves). All distributions are normalized. The two numbers in the top-right corner of each panel are the fractional parallax uncertainty σ_ϖ/ϖ (upper), and the parallax in mas (lower). Panels (a) to (i) are unimodal posteriors for which the vertical solid line is the mode and the vertical dashed lines denote the HDI containing 0.68 of the posterior probability. In panel (a) the prior is so broad on this scale that it appears at almost zero density. Panels (a) to (h) are ordered by increasing σ_ϖ/ϖ. Panel (i) is for a negative parallax. Panel (j) shows a bimodal solution for which the HDI cannot be computed: the solid line shows the median, and the dashed lines show the ETI containing approximately 0.68 of the posterior probability.

greater than 5 (7.7% greater than 10, 0.98% greater than 50). These negative and/or low SNR parallaxes are due to the numerical dominance of relatively distant and faint sources. Cumulative distributions for σ_ϖ/ϖ are shown in panels (c) and (d) for positive (black curve) and negative (red curve) parallaxes.

3.2. *Example distance posteriors*

Figure 5 shows examples of several posteriors and their corresponding priors. (More examples can be found in Figure 12 of Bailer-Jones 2015.) Recall that the mode of the posterior is at $2L_{\rm sph}$. We do not plot the likelihood because it is an improper distribution in r, so can only be plotted with an arbitrary scaling. The exact shape of the posterior depends in a somewhat complex manner on the value of all three of its variables: ϖ, σ_ϖ/ϖ, and $L_{\rm sph}$. These examples have been chosen to show a range of behaviours; they do not represent the frequency of such posteriors among the

results. Panels (a) to (c) are for small positive values of σ_ϖ/ϖ: these posteriors are dominated by the likelihood (as opposed to the prior) and are therefore approximately Gaussian with a mode close to the inverse parallax. Panels (d) and (e) are cases where the likelihood and prior have similar size influence on the posterior. As σ_ϖ/ϖ increases, the positive distance tail extends and the posterior becomes increasingly asymmetric. Panel (d) is a relatively rare example of a posterior with a high, narrow peak accompanied by a long tail.

Broadly speaking, the larger σ_ϖ/ϖ, the more the prior dominates the posterior. This is seen in panels (f) and (h). However, it is not simply the case that the larger the value of σ_ϖ/ϖ, the closer the posterior is to the prior. We see this in panel (g), where $\sigma_\varpi/\varpi = 2.64$ (larger than in panel f), yet the posterior mode is at more than twice the value of the prior mode. The posterior is always the (normalized) product of prior and likelihood, so if these peak at very different values, as is the case in panel (g) ($2L_{\rm sph} = 1.0\,$kpc vs. $1/\varpi = 10.4\,$kpc), the likelihood can still play a significant role, even when the data are poor (i.e. σ_ϖ/ϖ is large). In other words, prior dominance does not increase monotonically with σ_ϖ/ϖ. The actual values of the parallax and the prior length scale are also important.

Panel (i) is an example of a negative parallax. Panel (j) shows the relatively rare case of a bimodal posterior. In this particular case the HDI cannot be uniquely defined, so our algorithm returns the median as the point estimate, and the equal-tailed quantiles to define the asymmetric confidence interval.

3.3. *Distance statistics*

Figure 6 summarizes the properties of the distance estimates for unimodal posteriors with mode and HDI estimates. We discuss this figure for most of the rest of this section.

Panel (a) shows the distribution of the distance estimate $r_{\rm est}$, with the distribution of the length scale, $L_{\rm sph}$, in panel (d) beneath it for comparison. Some idea of the fractional distance uncertainties is shown by the histogram in panel (b), where $(r_{\rm hi} - r_{\rm lo})/2$ is taken as a single symmetrized measure of the uncertainty. This can be compared with the distribution of the fractional parallax uncertainties shown in Figure 4b. The fractional distance uncertainties are generally smaller because the prior prevents the posterior distribution becoming arbitrarily wide.

Panel (c) indicates the skewnesses of the posteriors. The upper part of the confidence interval ($r_{\rm hi} - r_{\rm est}$) is always larger than the lower part ($r_{\rm est} - r_{\rm lo}$). In 2%, 0.6%, and 0.15% of the cases it exceeds it by factors of more than 5, 10, and 20 respectively. As mentioned earlier, if the asymmetry is so large that the positive tail is nearly flat, then the HDI may not have been estimated very accurately.

Panel (e) shows how the fractional distance uncertainty varies with distance. The complex shape is a consequence of the true distribution in the Galaxy, the variation of $L_{\rm sph}(l,b)$, and the properties of the posterior.

The relation between the fractional parallax uncertainty and distance uncertainty is shown in panel (f). For small (positive) values of both parameters they are highly correlated. For large positive values of σ_ϖ/ϖ, the fractional distance uncertainty can never get very large, due to the prior. For the prior itself, $(r_{\rm hi} - r_{\rm lo})/(2r_{\rm mode}) \simeq 0.74$ (independent of L), and this is the value that the upper envelope of the distribution asymptotes to as $\sigma_\varpi/\varpi \to \infty$. For some intermediate values of σ_ϖ/ϖ the fractional distance uncertainties are quite large. These are due to posteriors with long positive tails like that shown in Figure 5(d): $r_{\rm hi} \gg r_{\rm est}$ which results in $(r_{\rm hi} - r_{\rm lo})/2r_{\rm est}$ being large. If we used the median or mean as a distance estimate instead, this measure of the fractional distance uncertainty would not extend to such large values (although those estimators would introduce other problems).

Panel (g) plots the ratio of the mode of the posterior ($r_{\rm est}$) to the mode of the prior ($2L_{\rm sph}$) as a function of the fractional parallax uncertainty, σ_ϖ/ϖ, to give some sense of how dominant the prior is. The posterior is a (renormalized) product of the prior and likelihood, so the location of the posterior mode depends not only on the modes of the prior and likelihood – the latter is at $1/\varpi$ when considered as a function of distance – but also on their widths, i.e. how peaked/flat the distributions are. This generates a rather complex behaviour in panel (g). For small positive values of σ_ϖ/ϖ (below 0.1–0.2) we see a wide range of values of $r_{\rm est}/2L_{\rm sph}$, as we would expect: there is little constraint from the prior. As σ_ϖ/ϖ increases, the prior plays more of a role and so the posterior mode deviates from it by less. We see a sharp lower boundary in the scatter plot, showing that $r_{\rm est}$ cannot drop below some finite multiple of the prior mode. This is due to the shape of the prior density function, which always drops to zero at zero distance: for a likelihood of finite width (i.e. $\sigma_\varpi/\varpi > 0$), its product with the prior always produces a mode above zero. The upper envelope, in contrast, is blurred because there is no hard upper limit on the location of the prior mode.

A comparison of the distances to the naive inverse parallax estimates is shown in panel (h) by plotting their ratio (i.e. $r_{\rm est}\varpi$), and zoomed in for the higher-precision parallaxes in panel (i). Once σ_ϖ/ϖ exceeds around 0.1–0.3, the inverse parallax deviates considerably from $r_{\rm est}$. For large fractional parallax uncertainty, $r_{\rm est}$ tends to be less than $1/\varpi$. This

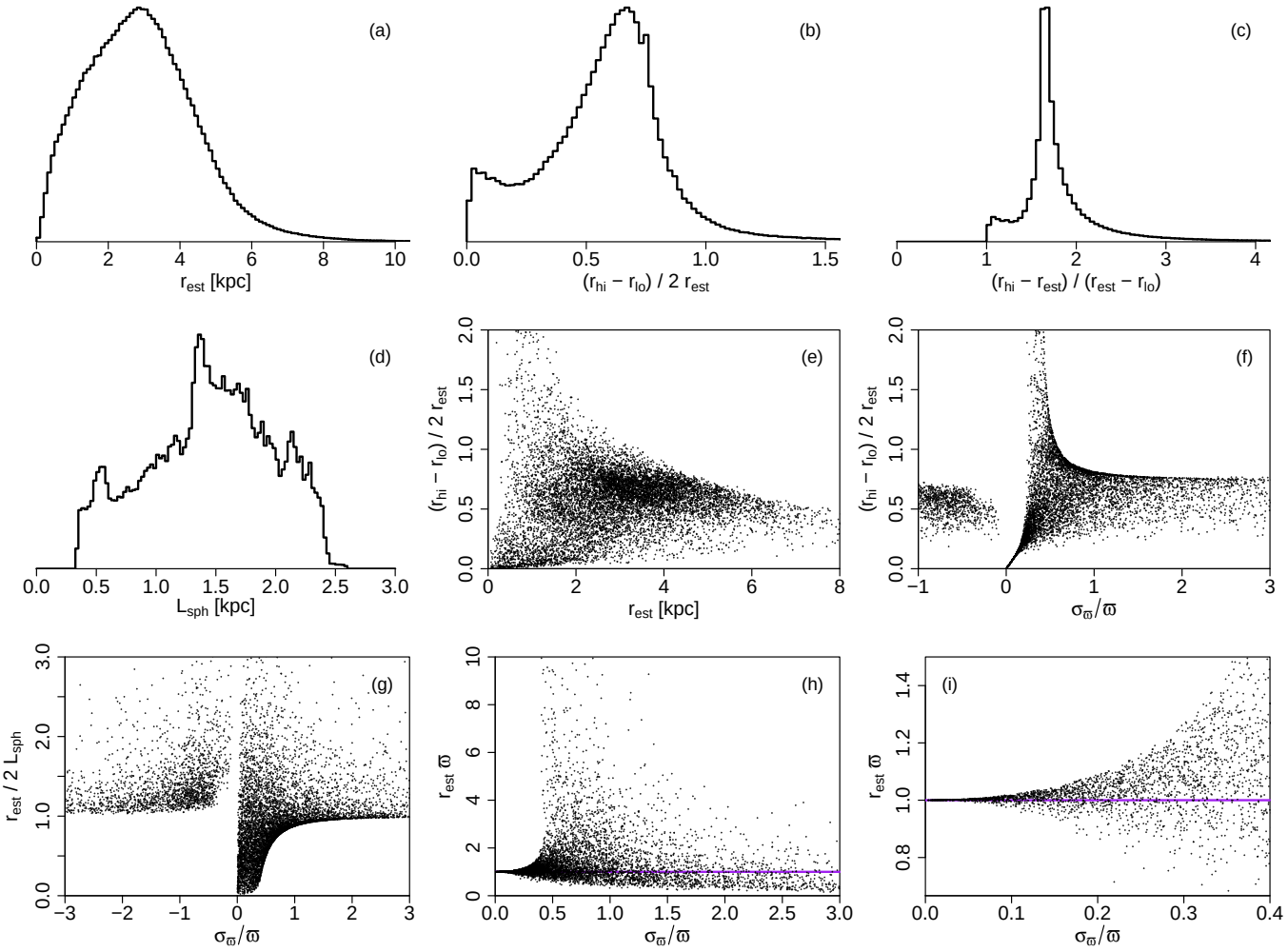

Figure 6. Summaries of the distance inference for a randomly selected set of one million sources with `result_flag=1` and `modality_flag=1`. Linear scales are used on both axes of the histograms. To avoid crowding, panels (e)–(i) plot just 10 000 of the sources. Note that panel (d) shows the distribution of $L_{\rm sph}$ over stars, whereas the histograms of L in Figures 1b and 2b are the density over equal-area cells on the sky. In panels (h) and (i) – the latter is a zoom of the former – the horizontal purple line indicates $r_{\rm est} = 1/\varpi$.

is because a noisy inverse parallax can extend to arbitrarily large distances, something that the prior constrains. The inverse parallax is known to be a biased and very noisy estimator (e.g. Bailer-Jones 2015; Luri et al. 2018) which tends to overestimates the true distance.

3.4. *Spatial distribution*

The left panel of Figure 7 shows the distribution of estimated distances in Galactic coordinates, for those stars where the posterior is unimodal and this mode is the distance estimator. Compare this to the right panel, which shows the distribution of the modes of the corresponding priors. The overall features of the prior remain evident in the left panel, partly because such features correspond to real, previously known features in the Galaxy, but also because most stars have large fractional parallax uncertainties. Yet we clearly see features that are not present in the prior map, such as the Large and Small Magellanic Clouds. Our mean distances to these satellite galaxies are of course underestimated, because the stars have very large fractional parallax uncertainties, so our estimates are prior-dominated. The largest length scales in the prior model are only about 2.6 kpc – corresponding to prior modes of 5.2 kpc – whereas these galaxies are in fact about 50–60 kpc away. The same problem applies to distant giants, because the prior will normally be dominated by the nearer dwarfs in the model.

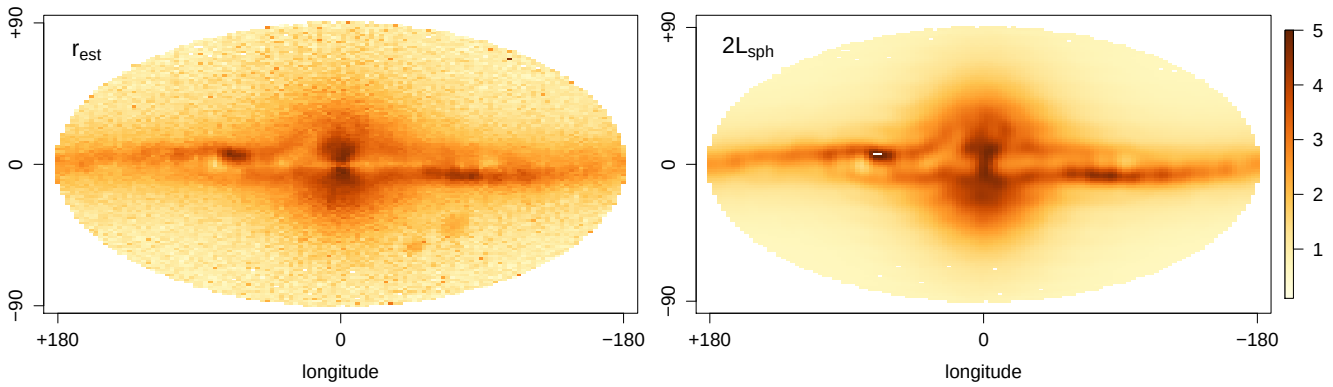

Figure 7. Distribution of distances for sources with `result_flag=1` and `modality_flag=1` shown in Galactic coordinates on a Mollweide projection. The colour scale (the same in both panels) shows the mean distance, in kpc on a linear scale, for sources over a small range of (l, b). White cells lie outside the range 0–5 kpc. Left: The estimated distances in the catalogue, $r_{\rm est}$ (i.e. the posterior mode). Right: The prior mode ($= 2L_{\rm sph}$).

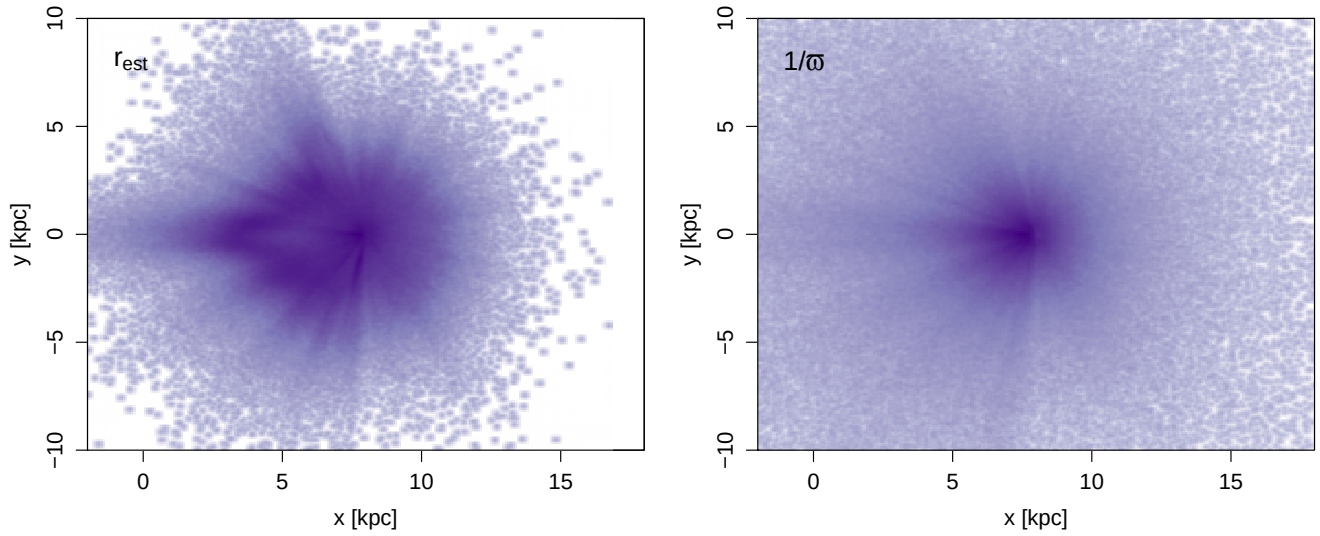

Figure 8. Distribution of stars' positions projected onto the Galactic plane shown as a density plot. The Galactic Centre is at $(0, 0)$ kpc and *Gaia* (and the Sun) are at $(+8, 0)$ kpc. The left panel is constructed using our distance estimates (for `result_flag=1` and `modality_flag=1`). The right panel is constructed using the naive $1/\varpi$ distance estimator (with the additional requirement that $\varpi > 0$).

Figure 8 compares the projected distribution of stars in the Galactic plane from our distance estimates (left) with those obtained from a naive parallax inversion (right). This shows the tendency of inverse parallax to extend to implausibly large distances. The radial lines from the Sun in both panels are due to nearby dust clouds diminishing the number of stars observed along certain sight-lines.

3.5. *Validation using clusters*

A simple validation of our distance estimates can be obtained using stellar clusters, on the assumption that the members of a physical (gravitationally bound) cluster span just a small range of distances.

Given a list of clusters with known sky positions, we select cluster members based on their proper motions. That is, we assume all members follow a common space motion (with some dispersion) that is different from the bulk of the field stars in the same region. We model the proper motion plane using Gaussian mixture models for both the cluster

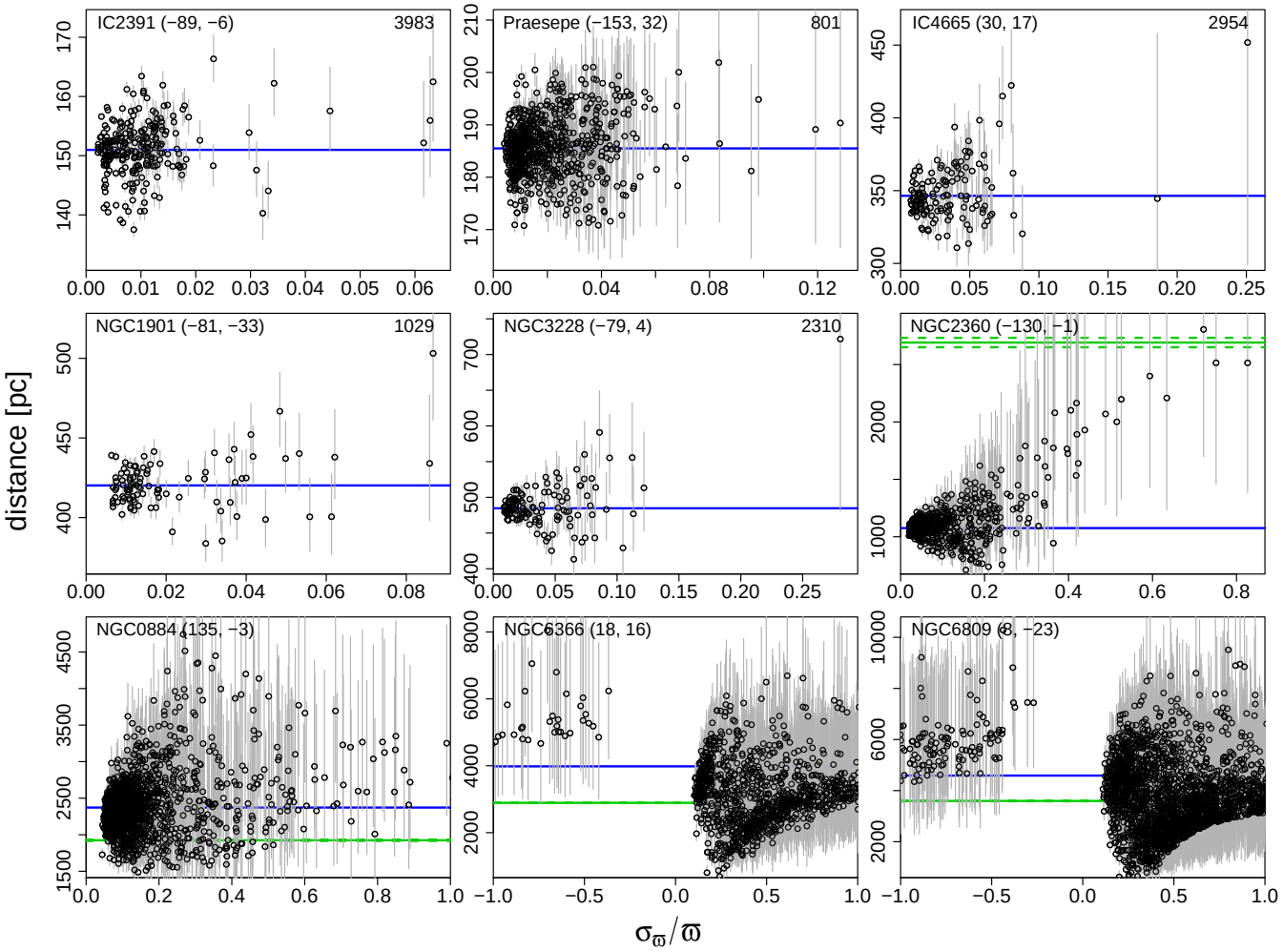

Figure 9. Validation of distance estimates using clusters (one per panel, with the name and (l, b) coordinates shown in the top left corner of each). We plot the estimated distances, $r_{\rm est}$, of the cluster members as open circles, as a function of σ_ϖ/ϖ. The error bars show the lower ($r_{\rm lo}$) and upper ($r_{\rm hi}$) bounds of the confidence interval for each star. The range of σ_ϖ/ϖ in each panel is from $\max(-1, \min(0, \sigma_\varpi/\varpi))$ to $\min(1, \max(\sigma_\varpi/\varpi))$. The vertical range of each panel is set to span the full range of $r_{\rm est}$ for everything lying in the range of σ_ϖ/ϖ plotted. The solid green horizontal line is $2\overline{L_{\rm sph}}$, the average of the modes of the priors for all cluster members. The dashed green lines around this (sometimes barely visible) show the standard deviation in this across the cluster members. In several cases these measures are outside the limit of the plot, in which case $2\overline{L_{\rm sph}}$ is written in the top right corner of the panel. The solid blue line is the inverse of the mean cluster parallax for all members (including any beyond the σ_ϖ/ϖ limits plotted). The clusters are ordered by increasing cluster distance.

region and a field region. This allows us to assess the membership probability to the cluster on a star-by-star basis. For present purposes we focused on purity by selecting only the most probable members. This method gives very a similar selection of members to that found and reported in Gaia Collaboration (2018b).

Figure 9 shows some results of validating our distance estimates using clusters. Each panel shows our distances to individual cluster members as a function of the fractional parallax uncertainty. (In all cases the estimates turn out to have `result_flag=1` and `modality_flag=1`.) These distances can be compared to the inverse of the mean parallax of the cluster members, shown as the blue line. This mean is computed using all members regardless of σ_ϖ/ϖ (including negative parallaxes if present). Although the parallax uncertainties are correlated on the angular scale of most stellar clusters ($\lesssim 1^\circ$), these correlations need not be considered when forming a simple mean. They would need to be accommodated if we estimated the formal uncertainty in this mean. Yet due to the large number of cluster members, the uncertainty in the mean is instead limited by the systematic parallax error, which is generally

Table 1. The format of the distance catalogue, showing data on five sources chosen at random. The source_id is a long integer, and is the same as that in GDR2. $r_{\rm est}$ is the point distance estimate. $r_{\rm lo}$ and $r_{\rm hi}$ are the lower and upper bounds of the (approximately) 68% confidence interval. $L_{\rm sph}$ is the length scale of the prior used. Distances are shown rounded to 0.001 pc here, but are given at full numerical precision in the catalogue. `result_flag` can take values: 0 (no solution, so $r_{\rm est}$, $r_{\rm lo}$, and $r_{\rm hi}$ are all `NaN`); 1 ($r_{\rm est}$ is the mode, and $r_{\rm lo}$ and $r_{\rm hi}$ define the highest density interval, HDI); 2 ($r_{\rm est}$ is the median, and $r_{\rm lo}$ and $r_{\rm hi}$ define the equal-tailed interval, ETI). `modality_flag` gives the number of modes of the posterior: it is either 1 or 2. If `result_flag=1` for bimodal posteriors, the mode reported is the highest mode, and the highest density interval is defined around this.

source_id	$r_{\rm est}$ pc	$r_{\rm lo}$ pc	$r_{\rm hi}$ pc	$L_{\rm sph}$ pc	result flag	modality flag
6056824904455202560	2471.511	1144.487	4870.877	1324.036	1	1
3125855551398349952	2914.102	1456.329	5343.421	1307.715	1	1
5545891471750684160	3672.499	2358.817	5960.471	1340.001	1	1
1824805611208672256	2919.521	1338.509	5546.599	1407.083	1	1
4663657669625864448	1026.809	583.985	1896.781	504.083	1	1

Table 2. Minimum and maximum values of the seven columns in the distance catalogue, as well as the 1st, 50th, and 99th quantiles of the four distance columns. For all sources $r_{\rm hi} > r_{\rm est} > r_{\rm lo}$. $r_{\rm lo}$ can be exactly zero. The smallest values of $r_{\rm est}$ are almost certainly the result of spurious parallaxes.

	source_id	r_est	r_lo	r_hi	r_len	result_flag	modality_flag
minimum	4295806720	0.540	0.000	0.540	335.227	0	1
1st percentile		318.91	238.373	397.475	384.145		
50th percentile		2841.107	1572.914	5193.845	1473.436		
99th percentile		7317.404	5074.212	11 169.098	2388.871		
maximum	6917528997577384320	32 978.685	27 577.405	155 560 896.912	2597.694	2	2
any `NaN`?	no	yes	yes	yes	no	no	no

below 0.1 mas (Lindegren et al. 2018). Thus for clusters within one kpc or so, the inverse mean parallax is a reasonable approximation of the cluster distance. We are not suggesting that this is the best way to estimate true cluster distances. We are just using it to give some baseline against which to compare our distance estimates (and uncertainties therein).

Figure 9 shows that for nearby clusters (first five panels), our distance estimates for the individual stars agree with the cluster mean. This is as expected, since for nearby clusters all the members have small σ_ϖ/ϖ, and for small σ_ϖ/ϖ the posterior mode is similar to inverse parallax. In all of these cases the prior has negligible impact (the modes of the priors are much larger than the estimated distances). These panels also show that the estimated uncertainties (the error bars plotted) are reasonable, considering that clusters have a finite depth.

As we move to more distant clusters, we see some informative behaviour. First, we see that some cluster members have large σ_ϖ/ϖ. As discussed earlier, the posteriors for stars with large σ_ϖ/ϖ are often dominated by the prior (e.g. Figure 5f). This is seen nicely in the panel for NGC2360, where the estimates move up towards the prior for large σ_ϖ/ϖ. The error bars also get larger. The next cluster, NGC0884, doesn't show such a clear trend, and in fact at large σ_ϖ/ϖ the distance estimates are systematically larger than the prior. This is an example where even for poor parallaxes the posterior is not entirely dominated by the prior, because the likelihood, even though broad, suggests quite a different distance than the prior does. This was also seen in Figure 5g. The final two clusters are dominated by stars with large σ_ϖ/ϖ. Here the posteriors tend to be closer to the priors. Both clusters are so distant that many of their members have negative parallaxes. Yet our inferred distances to these are broadly consistent with the mean cluster parallax. Large uncertainties notwithstanding, this demonstrates that useful information can be extracted from negative parallaxes. Of course for these more distant clusters, the mean cluster parallax has a significant systematic error (order 0.1 mas), as do the individual parallaxes used for our distance estimates.

4. DISTANCE CATALOGUE CONTENT AND CHARACTERISTICS

Table 3. The numbers of sources with different combinations of the flags. The total number of sources is 1 331 909 727. HDI = highest density interval; ETI = equal-tailed interval.

		modality flag		
		1	2	Sums
result flag	(fail) 0	3278	0	3278
	(mode+HDI) 1	1 330 715 981	295 451	1 331 011 432
	(median+ETI) 2	13 477	881 540	895 017
	Sums	1 330 732 736	1 176 991	

The distance catalogue comprises 1 331 909 727 entries with seven columns. An extract for five sources selected at random, together with the definition of the columns, is given in Table 1. The minimum and maximum values of the columns are given in Table 2. Table 3 summarizes the six different types of solutions possible, corresponding to the six combinations of the two flags. The vast majority of posteriors are unimodal and could be successfully summarized with the mode and HDI (`result_flag=1`, `modality_flag=1`). A very small fraction are instead summarized with the median and ETI (`result_flag=2`, `modality_flag=1`). Some bimodal posteriors are summarized with the mode and HDI (`result_flag=1`, `modality_flag=2`), although for most the HDI is not uniquely defined, so the median and ETI are reported instead (`result_flag=2`, `modality_flag=2`).

As noted before, in a very small fraction of cases the posterior is so skew that the computed HDI actually spans a probability that is much larger than 0.6827 (but not as high as 0.9, which is what forces the solution to `result_flag=2`). Such posteriors can be identified from the fact that then $r_{\rm hi} - r_{\rm est} \gg r_{\rm est} - r_{\rm lo}$.

When doing statistical analyses on an ensemble of the catalogue results one should not mix the two types of solution (values of `result_flag`), as the estimates are conceptually different. If one is interested in using the bimodal solutions, we strongly suggest one plots the posteriors and decide whether the summaries provided are appropriate. Note that adopting a different length scale could remove the bimodality.

No post-processing has been performed to modify distances or to remove entries which could potentially be identified as implausible based on additional information (such as a source classification, photometry, or proper motions). GDR2 is known to contain a number of sources with spuriously large parallaxes, some of which are (spuriously) precise (e.g. 174 sources with $\varpi > 500$ mas, all of which have $\sigma_\varpi/\varpi < 0.01$). This produces, for example, 21 sources in our distance catalogue with $r_{\rm est} < 1$ pc (they all have $G > 19$). While they could in principle be nearby brown dwarfs (and many have high proper motions), they are probably cross matching errors in the data processing. It is very unlikely that *Gaia* has discovered so many new nearby sources. More conservative filtering described in Appendix C of Lindegren et al. (2018) preferentially removes such sources, suggesting most are indeed spurious. Arenou et al. (2018) (section 4.1 and Figure 10) show that spuriously large parallaxes are generally concentrated in regions of high stellar density.

The probability density of our prior reaches zero only asymptotically at large distances, so in principle extragalactic objects can be assigned arbitrarily large distances corresponding to their very small parallaxes. In practice, however, the half million or so extragalactic objects in GDR2 have statistically insignificant parallaxes, so the prior will dominate their distance estimates, meaning that these will be woefully too small. The same applies for stars in the local group.

Just as the *Gaia* astrometric data processing treats all sources as single stars moving on linear paths relative to the Sun, so our distance estimates ignore any such binarity.

The individual distance estimates are coupled on angular scales on the sky through the use of a smooth model for the prior length scales, $L_{\rm sph}(l, b)$. This is important when the prior dominates the posterior, because the combination of multiple distance estimates (or even the entire posterior) for such stars will compound the prior. As described in Lindegren et al. (2018), the parallax uncertainties show between-source correlations on various angular scales on the sky, particularly below 1 degree. Our distance uncertainties will be similarly correlated. This must be considered when combining data, for example to estimate the distance to a cluster based on several of its members. We do not recommend using distance estimates in such combinations, however. It is better to make a model of the cluster and infer its distance and size simultaneously. A tutorial by one of us (CBJ) describing a simple approach to this can be found in Luri et al. (2018).

While it would have been convenient for the user if these distances were part of the main GDR2 catalogue, it was decided by the DPAC Executive not to include any distances in the official release. Our distance catalogue is nonetheless available on the Gaia archive http://gea.esac.esa.int/archive/ as an external catalogue table called `external.gaiadr2_geometric_distance`. It is therefore easy to make queries which join the two tables on the source identifier. The following ADQL query gives an example of how to do this by selecting nearby hot stars in a single healpix at level 4.

```
SELECT source_id, ra, dec, phot_g_mean_mag, r_est, r_lo, r_hi, teff_val
FROM external.gaiadr2_geometric_distance
JOIN gaiadr2.gaia_source USING (source_id)
WHERE r_est < 1000 AND teff_val > 7000
AND source_id < 2251799813685248
```

5. SUMMARY AND CONCLUSIONS

We have produced a catalogue of distance estimates for 1.33 billion sources in the second *Gaia* data release. By using a inference approach we can infer meaningful distances even when the parallaxes are negative and/or the parallax precision is low. These are in fact the most common cases in the *Gaia* catalogue. The estimation of the distance to each source involves a prior with a single length scale parameter. This parameter, obtained by fitting to a three-dimensional model of the Galaxy as seen by *Gaia*, varies smoothly and realistically as function of Galactic longitude and latitude (only). In the limit of precise parallaxes, say better than about 10% (77 million sources), the prior has little impact on the resulting distance estimates. As the parallax precision degrades, the information from the measurement becomes poorer, and our background information as represented by the prior starts to take over. The advantage of the inference approach is that as it ensures a smooth and graceful transition between the data-dominated and prior-dominated regimes, thus assuring self-consistency over the entire catalogue. This is important, because selecting sources with small fractional parallax uncertainties – perhaps in the hope of using inverse parallax as a distance estimator – will create a sample biased toward bright nearby sources.

The other advantage of the probabilistic approach adopted here is that we also infer a sensible confidence interval on every distance estimate. These must not be ignored, as there is a genuine and unavoidable uncertainty in the distance estimates. We specify this confidence interval via its lower and upper bounds. These are asymmetric with respect to the distance estimate on account of the nonlinear transformation from parallax to distance.

We remind the reader that it was our explicit intention to provide purely geometric distances, free from assumptions of stellar physics or interstellar extinction toward individual sources. We believe such distances and their uncertainties are useful for various cases such as those described in the introduction. We nonetheless realise that more precise – if not necessarily more accurate – distances can be estimated for a targeted subset of stars for which we use more information. Finally, we should emphasise that there are many tasks where explicit distances should not be used at all. If the distance is only an intermediate step in an analysis, for example when estimating luminosities, or transverse velocities from the proper motions, you will almost always want to do the inference from first principles, working directly with the parallaxes.

This work was funded in part by the DLR (German space agency) via grant 50 QG 1403. It has made use of data from the European Space Agency (ESA) mission *Gaia* (http://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, http://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. We thank the IT departments at MPIA and ARI for computing support.

REFERENCES

Anderson, L., Hogg, D. W., Leistedt, B., Price-Whelan, A. M., & Bovy, J. 2017, ArXiv e-prints, arXiv:1706.05055

Arenou, F., Luri, X., Babusiaux, C., et al. 2018, ArXiv e-prints, arXiv:1804.09375

Astraatmadja, T. L., & Bailer-Jones, C. A. L. 2016a, ApJ, 832, 137

Astraatmadja, T. L., & Bailer-Jones, C. A. L. 2016b, ApJ, 833, 119

Bailer-Jones, C. A. L. 2015, PASP, 127, 994

Bailer-Jones, C. A. L. 2017, Practical Bayesian Inference (Cambridge University Press)

Cantat-Gaudin, T., Vallenari, A., Sordo, R., et al. 2018, ArXiv e-prints, arXiv:1801.10042

Coronado, J., Rix, H.-W., & Trick, W. 2018, MNRAS, arXiv:1804.07760

Gaia Collaboration. 2018a, A&A, arXiv:1804.09365

—. 2018b, A&A, arXiv:1804.09378

Hyndman, R. J. 1996, The American Statistician, 50, 120

Leistedt, B., & Hogg, D. W. 2017, AJ, 154, 222

Lindegren, L., Lammers, U., Hobbs, D., et al. 2012, A&A, 538

Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, arXiv:1804.09366

Luri, X., Brown, A., Sarro, L., et al. 2018, A&A, arXiv:1804.09376

McMillan, P. J., Kordopatis, G., Kunder, A., et al. 2017, ArXiv e-prints, arXiv:1707.04554

O'Malley, E. M., Gilligan, C., & Chaboyer, B. 2017, ApJ, 838, 162

Queiroz, A. B. A., Anders, F., Santiago, B. X., et al. 2018, MNRAS, 476, 2556

Rybizki, J., Demleitner, M., Fouesneau, M., et al. 2018, ArXiv e-prints, arXiv:1804.01427

Schönrich, R., & Bergemann, M. 2014, MNRAS, 443, 698